Development of A Comprehensive Evaluation Scale for LLM-Powered Counseling Chatbots (CES-LCC) Using the eDelphi Method

Marco Bolpagni * and Silvia Gabrielli

Posted Date: 22 January 2025

doi: 10.20944/preprints202501.1621.v1

Keywords: counseling chatbots; mental health chatbots; large language models (LLMs); digital mental health; chatbot evaluation; eDelphi methodology; evaluation scale

Article

Development of A Comprehensive Evaluation Scale for LLM-Powered Counseling Chatbots (CES-LCC) Using the eDelphi Method

Marco Bolpagni [1,2,*] and Silvia Gabrielli [2]

[1] Human Inspired Technology Research Centre, University of Padova, 35121 Padova, Italy

[2] Digital Health Research, Centre for Digital Health and Wellbeing, Fondazione Bruno Kessler, 38123 Trento, Italy

* Correspondence: mbolpagni@fbk.eu

Abstract: Background/Objectives: With advancements in Large Language Models (LLMs), counseling chatbots are becoming vital tools for delivering scalable and accessible mental health support. Traditional evaluation scales, however, fail to adequately capture the sophisticated capabilities of these systems, such as personalized interactions, empathetic responses, and memory retention. This study aims to design a robust and comprehensive evaluation scale, the Comprehensive Evaluation Scale for LLM-Powered Counseling Chatbots (CES-LCC), using the eDelphi method to address this gap. **Methods**: A panel of 16 experts in psychology, artificial intelligence, human-computer interaction, and digital therapeutics participated in two iterative eDelphi rounds. The process focused on refining dimensions and items based on qualitative and quantitative feedback. Initial validation, conducted after assembling the final version of the scale, involved 49 participants using the CES-LCC to evaluate an LLM-powered chatbot delivering Self-Help Plus (SH+), an Acceptance and Commitment Therapy-based intervention for stress management.**Results**: The final version of the CES-LCC features 27 items grouped into nine dimensions: Understanding Requests, Providing Helpful Information, Clarity and Relevance of Responses, Language Quality, Trust, Emotional Support, Guidance and Direction, Memory, and Overall Satisfaction. Initial real-world validation revealed high internal consistency (Cronbach's alpha = 0.94), although minor adjustments are required for specific dimensions, such as Clarity and Relevance of Responses. **Conclusions:** The CES-LCC fills a critical gap in the evaluation of LLM-powered counseling chatbots, offering a standardized tool for assessing their multifaceted capabilities. While preliminary results are promising, further research is needed to validate the scale across diverse populations and settings.

Keywords: counseling chatbots; mental health chatbots; large language models (LLMs); digital mental health; chatbot evaluation; eDelphi methodology; evaluation scale

1. Introduction

1.1. Background

Counseling chatbots are conversational agents designed to provide mental health support, guidance, and therapeutic conversations [1]. These chatbots simulate human-like interactions to help individuals manage their emotional well-being, particularly in situations where immediate access to human counselors is unavailable. By offering scalable and accessible mental health services, counseling chatbots address key barriers such as cost and geographical limitations, making them an appealing solution for both users and healthcare systems [2]. The adoption of counseling chatbots has accelerated due to increasing mental health awareness and the growing demand for scalable solutions, a trend that has been further amplified by the COVID-19 pandemic [3]. As mental health

 doi:10.20944/preprints202501.1621.v1

issues become more pressing globally, the role of these technologies is expected to expand in the next years. Traditionally, counseling chatbots relied on rule-based systems and Natural Language Processing (NLP) techniques to interpret and respond to user inputs [4]. However, recent advancements in Large Language Models (LLMs), such as GPT-3 and GPT-4, represent a significant upgrade. LLMs have vastly improved the quality of text generated, the level of personalization, and the contextual awareness of chatbot interactions, enabling more nuanced and emotionally resonant conversations [5]. As counseling chatbots become more sophisticated due to the introduction of LLMs, so must the methods used to evaluate them. Traditional evaluation scales, centered on user satisfaction and related concepts, are valuable but may no longer suffice in capturing the full spectrum of these systems' capabilities. The integration of LLMs introduces new aspects, such as enhanced empathy, seamless conversational flow, and contextually appropriate emotional responses, which require more nuanced and multifaceted evaluation tools.

1.2. Related Works

The evaluation of mental health chatbots is a complex task that involves assessing multiple aspects of their performance and impact on users. Effective mental health chatbots must engage users, provide a positive user experience, be easy to use, offer helpful and empathetic support, foster trust and alliance, and demonstrate strong technical performance in terms of language quality [6]. As such, researchers and developers have identified several key aspects that are crucial to the success of these systems. Engagement [7] and user experience [8], for example, can influence users' motivation to continue using the chatbot and their overall satisfaction with the system. Usability [9] is critical to ensuring that users can effectively interact with the chatbot and access the support they need. Perceived helpfulness, empathy, trust, support, and alliance are all essential components of a therapeutic relationship and are critical to establishing a sense of rapport and connection between the user and the chatbot [6]. Technical performance and language quality, meanwhile, are fundamental to ensure that the chatbot can provide accurate and informative responses to users' queries.

To evaluate these dimensions, researchers have employed a range of scales and metrics. Engagement, for instance, has been evaluated [10], [11] using the User Engagement Scale (UES) [12]. This scale provides insights into users' emotional and cognitive investment in interacting with the chatbot. In contrast, user experience has been assessed [13], [14], [15] through the User Experience Questionnaire (UEQ) [16], which captures users' subjective experience of using the chatbot in terms of attractiveness, perspicuity, efficiency, dependability, stimulation and novelty. To evaluate usability, instead, researchers have utilized [17], [18], [19], [20] several scales, including the System Usability Scale (SUS) [21], Chatbot Usability Questionnaire (CUQ) [22], and Bot Usability Scale (BUS) [23]. These scales provide a comprehensive understanding of users' perceptions of the chatbot's ease of use. Perceived helpfulness, which refers to users' beliefs about the chatbot's ability to provide effective support, has been evaluated [24], [25] using frameworks such as the Unified Theory of Acceptance and Use of Technology (UTAUT) [26] and Perceived Usefulness and Ease of Use (PEOU) [27]. These frameworks help researchers understand the factors that influence users' intentions to use mental health chatbots. Empathy, a critical component of human-computer interaction in mental health contexts, has been assessed [28], [29] using scales such as the Perceived Empathy of Technology Scale (PETS) [30] and Empathy Scale for Human–Computer Communication (ESHCC) [31]. These scales capture users' perceptions of the chatbot's ability to understand and respond to their emotional needs. In addition to empathy, perceived trust, support, and alliance are essential aspects of mental health chatbots. The Virtual Therapist Alliance Scale (VTAS) [32] has been used [33] to evaluate these dimensions, providing insights into users' perceptions of the chatbot as a supportive and trustworthy therapeutic agent. Technical performance and language quality instead are often evaluated [34], [35], [36], [37] using automated metrics such as perplexity, BLEU [38], and ROUGE [39]. These metrics provide quantitative insights into the chatbot's ability to generate coherent, contextually appropriate, and grammatically accurate responses.

 doi:10.20944/preprints202501.1621.v1

Despite the availability of these tools, many studies still rely on custom evaluation grids tailored to their research needs [40], [41], introducing variability across studies. These grids are often designed to cover all relevant aspects in a compact form, as using a full set of scales for the evaluations would be too lengthy and impractical for many studies. This lack of standardization, however, can hinder comparisons between studies and limit the generalizability of findings. As LLM-powered counseling chatbots become increasingly sophisticated, an integrated evaluation approach is essential to ensure that all aspects are adequately assessed while keeping the length of the scale manageable.

1.3. Aim

To address the limitations of current evaluation methods and provide a compact comprehensive tool designed for the unique demands of LLM-powered counseling chatbots, we aim to develop a novel scale (Comprehensive Evaluation Scale for LLM-Powered Counseling Chatbots (CES-LCC)) using the eDelphi method [42]. This approach is particularly well-suited to emerging fields like AI-driven counseling, where expert knowledge is still evolving and consensus on best practices has not yet been fully established. The eDelphi method facilitates online expert collaboration through multiple rounds of feedback, ensuring that key evaluation criteria are identified and refined iteratively. This structured, consensus-driven process allows for the development of a robust, adaptable evaluation tool that reflects the diverse, complex requirements of LLM-powered counseling chatbots.

2. Materials and Methods

2.1. Participants

Following the Delphi methodology [43], this study employed purposive sampling to assemble a panel of experts in counseling psychology, Artificial Intelligence (AI), and Human-Computer Interaction (HCI). The Delphi method does not mandate a statistically representative sample, thereby affording flexibility in panel size. Conventionally, panel sizes range between 15 to 30 participants [44], [45]. Given this flexibility, invitations were extended to 22 experts, selected for their professional experience with counseling technologies, LLM-powered systems, and chatbot development. Of these, 16 experts agreed to participate. Eligibility criteria included a minimum of three years of professional experience, relevant academic publications or industry contributions, and familiarity with LLM technologies. Recruitment took place in a single round between 26th June 2024 and 4th July 2024. Invitations were distributed via email, outlining the study's objectives and the participants' role in defining and refining the evaluation scale. Participation was voluntary, with an estimated time commitment of 30 minutes per round. Experts were given a two-week time window to complete the first-round survey, with a reminder sent at the halfway point to encourage response. Following the completion of the first round, a second round of the eDelphi process was conducted to refine and consolidate the initial feedback. Participants were presented with a summary of the first-round results, including aggregated ratings and qualitative comments, and were invited to reassess their responses based on the group's collective insights. The second-round survey was distributed on 20th July 2024, with a two-week response window and similar reminders to encourage participation.

2.1. Procedure

For our eDelphi study, we followed the four steps proposed by [46], which consist of (1) a preparatory phase, (2) eDelphi rounds, (3) data processing and analysis, and (4) conclusion and reporting. The procedure for the eDelphi study is visually summarized in Figure 1, which provides an overview of the steps and corresponding activities carried out at each stage.

 doi:10.20944/preprints202501.1621.v1

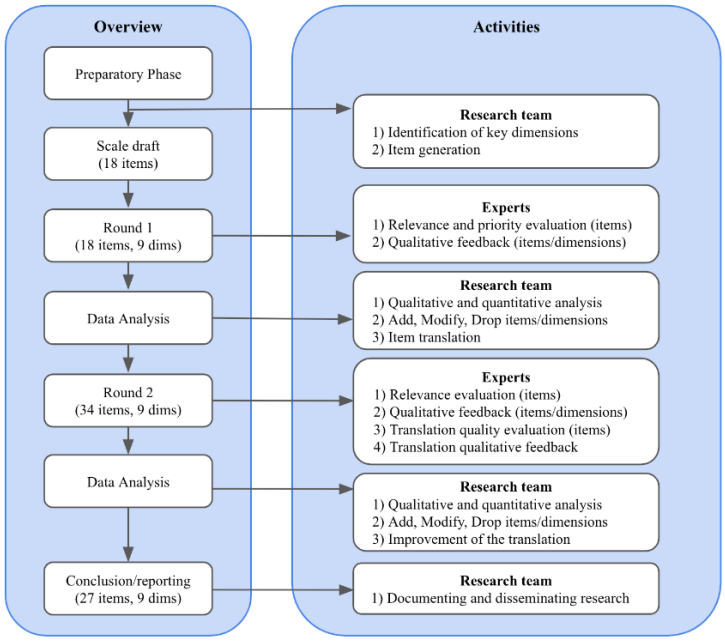

Figure 1. Overview of the process and activities carried on during the eDelphi study.

2.2.1. Preparatory Phase

The development of CES-LCC began with the identification of key dimensions critical to the assessment of LLM-powered counseling chatbots. Given that LLM-powered counseling chatbots represent a relatively new and rapidly evolving field, we adopted a targeted approach to the literature. Instead of conducting a comprehensive review across a broad range of sources, we focused on key publications and resources that specifically address chatbot evaluation in the context of mental health [40], [41], [47], [48] and LLM technologies [49], [50], [51], [52] as well as on the scales described in section 1.2. This targeted analysis allowed us to concentrate on the most relevant aspects to be evaluated, leading to the identification of recurring themes reflecting the technical, emotional, and linguistic dimensions relevant for the evaluation of LLM-powered counseling chatbots.

The initial pool of items was generated by M.B. and S.G., who have interdisciplinary expertise in AI, digital health system design, and psychology. The dimensions identified for the first draft of the scale included: Understanding my Requests, Providing Helpful Information, Clarity and Relevance of Responses, Ease of Use and Interaction, Language Quality, Trust, Emotional Support, Guidance and Direction, and Overall Satisfaction. Based on these dimensions, a total of 18 items (see Appendix A.1) were generated, with 2 items allocated to each category. We deliberately chose to limit the number of items per dimension to 2 to avoid guiding the experts too heavily and to allow for more open-ended feedback during the Delphi process.

2.2.2. eDelphi Rounds

The eDelphi process was structured into two iterative rounds aimed at refining and validating the scale. Both rounds were administered electronically via the Qualtrics XM platform [53].

First Round

During the initial round, experts were provided with the preliminary version of the scale comprising the eighteen items developed in the preparatory phase. Participants assessed each item's relevance using a five-point Likert scale ranging from “Not relevant at all” to “Very relevant” and its priority on a separate five-point Likert scale from “Very low” to “Very high.” Additionally, they offered qualitative feedback regarding each item's clarity and comprehensiveness, proposed additional items or dimensions. Following the first round, the original English scale was translated into Italian to ensure accessibility and applicability across both local and international contexts. The

 doi:10.20944/preprints202501.1621.v1

translation process followed a rigorous two-step approach to ensure linguistic and conceptual accuracy. Initially, the original scale (in English) was translated into Italian using a multilingual LLM, specifically Mistral Large [54]. Subsequently, two independent bilingual translators, both with expertise in artificial intelligence and psychology, undertook the refinement process. Each translator worked independently, leveraging their specialized knowledge to enhance the linguistic precision and conceptual clarity of the translation. Finally, the translators collaborated to consolidate their refinements into a cohesive and accurate final version.

Second Round

The revised scale, incorporating modifications from the first round, was inspected in the second eDelphi round by the same panel of experts. In this round, participants reviewed the updated scale, which included retained, revised, and newly added items or dimensions. Experts re-evaluated each item's relevance using the same five-point Likert scale and assessed priority by ranking the items in order of importance. This ranking method facilitated a more precise determination of each item's relative significance, helping future efforts to develop shorter versions of the scale by identifying the most critical items. Additionally, experts were asked to flag any items they deemed redundant and, if flagged, to specify which other item(s) the redundant item overlapped with. Participants also provided additional qualitative feedback to confirm whether the revisions effectively addressed prior concerns and were encouraged to suggest further enhancements or provide specific recommendations for improvement. Experts whose mother tongue is Italian were also asked to evaluate the translation quality, using both a five-point Likert scale from "Very poor" to "Excellent" and responding to the open-ended question: “Is there a better way to translate this item into Italian? If so, please provide the improved version below.”

2.2.3. Data Processing and Analysis

Throughout both eDelphi rounds, data processing involved detailed quantitative and qualitative analyses. For the first round, descriptive statistics (mean, median, interquartile range, and standard deviation) were calculated for both relevance and priority ratings of each item. In contrast, priority assessment in the second round was approached differently, utilizing rankings derived through the Borda count method [55]. This method aggregated participant rankings by assigning points inversely proportional to rank positions, providing a more structured framework for determining the collective prioritization of items. As in [46] items were retained if over 75% of participants rated them as 4 or 5 in relevance and if the interquartile range was below 2. Items with a mean relevance score below 3 or those failing to meet agreement criteria were excluded, along with any dimensions devoid of remaining items post-removal. Qualitative feedback from both rounds underwent thematic analysis [56] to extract common themes and suggestions related to item clarity, comprehensiveness, and potential oversights. This analysis was performed independently by M.B. and S.G. and subsequently consolidated through consensus. The insights derived from this analysis informed the necessary revisions and additions to the scale, thereby enhancing its overall quality and comprehensiveness. In the second round, an additional analysis was also performed to address redundancy among items. Since no established guidelines for redundancy were found in the literature, a statistical criterion was applied. Items flagged as redundant were analyzed, and those exceeding the third quartile of redundancy flags were systematically removed. This redundancy-focused refinement ensured a more concise and efficient evaluation scale, aligned with expert consensus.

To systematically guide and organize the refinement of the evaluation scale, we introduced the Add, Modify, Drop (AMD) approach, as summarized in the form of an algorithm in Figure 2. This framework consolidates established methods for decision-making regarding items/dimensions in the context of the development of new scales using the Delphi method [57], [58], [59], [60]. For each dimension, items were dropped based on the aforementioned quantitative criteria, new items were

 doi:10.20944/preprints202501.1621.v1

added in response to qualitative feedback to address identified gaps, and existing items were modified according to qualitative feedback to improve clarity, comprehensiveness, and relevance.

Require: D: Dimensions, I_d: Items in d, Quantitative thresholds: t_{agree}, t_{IQR}, t_{mean}, $t_{\text{redundancy}}$: Redundancy threshold, Q: Qualitative feedback.
Ensure: Refined scale with updated dimensions and items.
Initialize $D_{\text{refined}} \leftarrow D$, $I_d^{\text{refined}} \leftarrow I_d$ for all $d \in D$.
for each $d \in D_{\text{refined}}$ **do**
  for each $i \in I_d^{\text{refined}}$ **do**
    Compute: % agree, mean, IQR.
    if % $\geq t_{\text{agree}}$ **and** mean $> t_{\text{mean}}$ **and** IQR $< t_{\text{IQR}}$ **then**
      Retain i.
    else
      Drop i.
    end if
  end for
  Refine I_d^{refined} with Q: improve clarity, address gaps, and add missing items.
  Perform redundancy analysis:
  for each $i \in I_d^{\text{refined}}$ **do**
    Compute the number of flags for i based on redundancy checks.
    if flags for $i \geq t_{\text{redundancy}}$ **then**
      Remove i from I_d^{refined}.
    end if
  end for
  if $I_d^{\text{refined}} = \emptyset$ **then**
    Remove d from D_{refined}.
  end if
end for
return D_{refined} with updated I_d^{refined}.

Figure 2. Add, Modify, Drop (ADM) algorithm.

2.2.4. Conclusion and Reporting

Upon completing the two eDelphi rounds, the final version of CES-LCC was assembled, documented and prepared for dissemination (a full version of the scale is provided in Section 3.3).

2.3. Initial Validation in Real-World

To assess the reliability of CES-LCC, we conducted an initial validation in a real-world setting. This stage focused on testing the internal consistency of the scale items both globally and across its dimensions. Data collection involved 49 users (participants details in Appendix C.1) engaging with a LLM powered chatbot that delivered the first session of the Self-Help Plus (SH+), an Acceptance and Commitment Therapy (ACT) based intervention for stress management and prevention originally developed by the World Health Organization (WHO) [61]. In this session participants have been introduced to the chatbot, received information about stress, emotional storms as well as some exercises people can use to manage these situations (e.g. grounding, focused attention). Participants filled the CES-LCC after a single interaction with the chatbot. Items were rated on a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". The reliability of the evaluation scale was assessed by calculating Cronbach's alpha [62] for each of the nine dimensions, as well as for the overall scale. Additionally, both item-total correlations [63] and inter-item correlations were examined for each dimension to verify that individual items contribute meaningfully to their respective constructs without introducing strong redundancy.

3. Results

3.1. Demographic Description of Experts

The expert panel consisted of 16 professionals with expertise in psychology, AI, HCI, and digital therapeutics (DTx) (Table 1). Gender representation was balanced, with 56.25% male and 43.75% female participants, and academic qualifications were predominantly at the master's (50%) and doctoral (31.25%) levels. Participants' average age was 34.5 years (SD = 10.66), indicating moderate

 doi:10.20944/preprints202501.1621.v1

age variability, while professional roles ranged from researchers (50%) to AI developers (37.5%) and psychologists (12.5%). The group included a broad spectrum of professional experience, with half of the participants having junior roles (3–5 years in the field) and others contributing more senior level expertise (18.75% with 21+ years of experience). All participants were based in Italy. While the homogeneous geographical location offers cultural homogeneity at the same time it might limit the views offered by the participants involved (more is discussed in Section 4).

Table 1. Demographic and professional characteristics of the expert panel.

Characteristic		**Value or % (n)**
Age		M = 34.50 (SD = 10.66)
Gender	Male	56.25% (9)
	Female	43.75% (7)
Education	Bachelor	12.50% (2)
	Master	50.00% (8)
	Doctorate	31.25% (5)
	PsyD Specialization	6.25% (1)
Area of expertise	Psychology	31.25% (5)
	Artificial Intelligence	31.25% (5)
	Human Computer Interaction	18.75% (3)
	Digital Therapeutics	18.75% (3)
Occupation	Researcher	50.00% (8)
	Developer (AI)	37.50% (6)
	Psychologist	12.50% (2)
Job seniority	3-5 years	50.00% (8)
	6-10 years	18.75% (3)
	11-15 years	12.50% (2)
	16-20 years	0.00% (0)
	21+ years	18.75% (3)
Country	Italy	100% (16)

3.2. First Round

During the first round of the eDelphi process, a thorough assessment of the original 18-item scale was undertaken. Item-level analysis revealed that 6 items did not meet the predetermined agreement criteria based on relevance and interquartile range (IQR) thresholds, leading to their exclusion from the scale (see Appendix A.1). Concurrently, experts provided qualitative feedback that resulted in the addition of 12 new items, enhancing the scale's ability to capture significant aspects that need to be evaluated in LLM-powered counseling chatbots. Additionally, five existing items were split into separate subitems to more effectively address distinct aspects, as some items were initially found to cover multiple, overlapping areas. To improve clarity, two items were rephrased based on the qualitative suggestions provided (see Appendix A.3). The dimension "Ease of use and interaction" was removed due to the absence of remaining items after the exclusion process. Meanwhile, a new dimension called "Memory" was introduced to evaluate the chatbot's ability to retain and utilize prior interactions effectively.

Experts recommended (see Appendix A.2) that each dimension should contain at least three items to meet the psychometric requirement for assessing internal consistency and to ensure the scale's reliability ("*Psychometrically, factors should have at least 3 items to be considered reliable, with 2 items it is not even possible to calculate internal consistency*"). Consequently, dimensions with fewer than three items after initial revisions were added new items to meet this requirement. Moreover, qualitative feedback highlighted the necessity of assessing privacy and security concerns related to the chatbot being assessed. However, experts concluded that these aspects pertain more to production nonfunctional requirements rather than intrinsic characteristics of the chatbot's functionality and, therefore, were excluded from the evaluation scale (*"Items related to data privacy and security might be relevant in this scenario. However, in my experience, these items are more aligned with production or implementation processes and might be better addressed under regulations like the EU AI Act and GDPR."*). Moreover, in relation to the "Trust" dimension, experts raised concerns about anthropomorphizing

 doi:10.20944/preprints202501.1621.v1

chatbots. They cautioned that attributing human-like qualities could skew the assessment of trustworthiness. As a result, all items containing any form of anthropomorphism were removed and replaced with new items proposed by the experts to better align with the construct while avoiding any reference to human characteristics. The priority assessment of the items did not provide a clear picture, as priority values ranged from 3.13 to 4.88, with an average value of 3.88 (SD = 0.41).

3.3. Second Round

The second round of the eDelphi process focused on refining the revised evaluation scale, which at this point consisted of 34 items across nine dimensions, incorporating feedback provided during the first round. General feedback from experts gathered in the second round indicated that the scale was comprehensive and complete. However, experts also remarked the presence of some redundancy in the scale, with certain items overlapping or duplicating information. The average relevance score of the items in this round was 4.32 (SD = 0.40), reflecting strong agreement among experts regarding the importance of the included items. Despite this general consensus, five items were removed due to low agreement, as they failed to meet the thresholds for relevance and IQR established in the methodology. Additionally, two items were excluded based on redundancy. The Italian translation of the scale received high ratings for quality, with an average score of 4.59 (SD = 0.31). However, comments from mother tongue experts (n = 12) highlighted the need for minor adjustments to improve linguistic precision and conceptual clarity. Consequently, 14 items were revised to enhance their translation quality and to maintain consistency between the English and Italian versions. Furthermore, four items were rephrased in both languages based on qualitative feedback. To complement these refinements, the collective ranking of the items for each dimension was computed. This analysis not only provides a foundation for the potential development of shorter versions of the scale, which may improve its deployment in practical application settings, but also makes it possible to identify the most relevant item within each dimension. These rankings, along with the finalized version of CES-LCC (27 items across 9 dimensions), are included in Table 2.

Table 2. Final version of CES-LCC. Items in italic are the Italian version.

Dimension	Item	Priority
Understanding requests [UR]	The chatbot consistently understands what I am saying and asking. *Il chatbot capisce ciò che sto dicendo e chiedendo.*	1
	The chatbot is able to make adequate inferences based on my messages. *Il chatbot è in grado di fare deduzioni appropriate basandosi sui miei messaggi.*	2
	The chatbot asks specific questions to better understand my requests. *Il chatbot fa domande specifiche per capire meglio le mie richieste.*	3
Providing helpful information [PHI]	The chatbot provides accurate information. *Il chatbot fornisce informazioni accurate.*	1
	The chatbot provides helpful information. *Il chatbot fornisce informazioni utili.*	2
	The chatbot provides information grounded in theory and scientific literature. *Il chatbot fornisce informazioni supportate da teorie e letteratura scientifica.*	3
Clarity and relevance of responses [CRR]	The chatbot's responses are clear, and easy to understand. *Le risposte del chatbot sono chiare e semplici da capire.*	1
	The chatbot's responses are adequately concise. *Le risposte del chatbot sono sufficientemente concise.*	2
	The chatbot's responses are irrelevant to my questions. *Le risposte del chatbot non sono pertinenti alle mie domande.*	3
Language quality [LQ]	The chatbot uses correct grammar and spelling in its responses. *Il chatbot fornisce risposte grammaticalmente e ortograficamente corrette.*	1
	The chatbot's language is appropriate for the context. *Il linguaggio del chatbot è appropriato al contesto.*	2
	The chatbot's language style sounds natural *Lo stile linguistico del chatbot suona naturale.*	3
Trust [T]	I feel safe sharing my personal matters with the chatbot. *Mi sento al sicuro nel condividere questioni personali con il chatbot.*	1
	I believe that the feedback and the information provided by the chatbot are trustworthy. *Credo che i feedback e le informazioni fornite dal chatbot siano affidabili.*	2

 doi:10.20944/preprints202501.1621.v1

	I believe the chatbot is transparent about its limitations and capabilities. *Credo che il chatbot sia trasparente riguardo ai suoi limiti e alle sue capacità*	3
Emotional support [ES]	The chatbot makes me feel heard and understood. *Il chatbot mi fa sentire ascoltato e capito.*	1
	The chatbot's responses feel empathetic and supportive. *Le risposte del chatbot risultano empatiche e supportive.*	2
	The chatbot's responses can make me feel reassured *Le risposte del chatbot sono in grado di farmi sentire rassicurato.*	3
Guidance and direction [GD]	The chatbot provides adjusted guidance in coping with my problems. *Il chatbot fornisce indicazioni personalizzate per aiutarmi a gestire i miei problemi.*	1
	The chatbot encourages me to take positive steps. *Il chatbot mi incoraggia a compiere azioni costruttive.*	2
	The chatbot helps me set realistic and achievable goals. *Il chatbot mi aiuta a stabilire obiettivi realistici e raggiungibili.*	3
Memory [M]	The chatbot accurately recalls details from previous conversations. *Il chatbot ricorda accuratamente i dettagli delle conversazioni precedenti.*	1
	The chatbot maintains consistency by integrating past interactions into current responses. *Il chatbot integra coerentemente le interazioni passate nelle risposte.*	2
	The chatbot adapts its advice based on information provided in earlier sessions. *Il chatbot adatta i suoi consigli in base alle informazioni fornite nelle sessioni precedenti.*	3
Overall satisfaction [OS]	I am overall satisfied with the usability of this chatbot. *Nel complesso, sono soddisfatto dell'usabilità di questo chatbot.*	1
	Overall, I feel that my interactions with the chatbot were worthwhile. *Nel complesso, trovo che le mie interazioni con il chatbot siano state proficue.*	2
	I am overall satisfied with the support provided by this chatbot *Nel complesso, sono soddisfatto del supporto offerto da questo chatbot.*	3

3.4. Initial Validation

The initial validation of the scale in a real-world setting demonstrated its reliability both globally and across individual dimensions. A total of 49 participants completed the evaluation scale after interacting with the LLM-powered chatbot that delivers the first session of the Self-Help Plus (SH+) intervention. The overall scale exhibited excellent internal consistency, with a Cronbach's alpha of 0.94, exceeding the generally accepted threshold of 0.70 for reliability [64]. Across the nine dimensions, Cronbach's alpha values ranged from 0.47 to 0.91, with 8 out of 9 dimensions exceeding the 0.70 threshold (see Table 3). These results suggest strong consistency for all the dimensions except for CRR (Clarity and Relevance of Responses), which reached only a Cronbach's alpha value of 0.47. Further investigation revealed that CRR3, the only reverse-coded item on the scale, was a key contributor to the lower reliability. Despite being recoded for the computation of Cronbach's alpha, it may have introduced additional cognitive complexity for respondents, potentially affecting the consistency of responses within this dimension. Average item-total correlations across the dimensions ranged from 0.33 to 0.82, with most dimensions showing satisfactory alignment between items and their respective constructs. The Emotional Support (ES) and Overall Satisfaction (OS) dimensions achieved the highest item-total correlations, with means of 0.82 and 0.80, respectively, reflecting strong coherence within these constructs. By contrast, the CRR dimension showed the lowest mean item-total correlation at 0.33, further highlighting the misalignment of CRR3 with the rest of the items in this dimension. Inter-item correlations provided additional insights into the internal structure of the scale. Mean inter-item correlations ranged from 0.28 (CRR) to 0.77 (ES). While most dimensions demonstrated inter-item correlations within the acceptable range (0.20–0.70) [65], ES and OS displayed notably high mean inter-item correlations (0.77 and 0.75, respectively), suggesting a need to investigate potential residual redundancies in these dimensions. These preliminary findings indicate that while the scale overall and most of its dimensions demonstrate acceptable psychometric properties, specific dimensions, such as CRR need further investigation.

 doi:10.20944/preprints202501.1621.v1

Table 3. Summary of inter-item correlations, item-total correlations, and Cronbach's α values for all the dimensions of the scale.

Dimension	Inter-Item Correlation		Item-total correlation		Cronbach's α
	Mean	Range	Mean	Range	
UR	0.42	0.28-0.54	0.50	0.41-0.61	0.68
PHI	0.58	0.44-0.73	0.65	0.55-0.76	0.79
CRR	0.28	0.02-0.71	0.33	0.06-0.54	0.47
LQ	0.40	0.23-0.61	0.47	0.31-0.63	0.63
T	0.55	0.41-0.74	0.63	0.49-0.73	0.78
ES	0.77	0.70-0.86	0.82	0.76-0.88	0.91
GD	0.48	0.33-0.73	0.54	0.38-0.64	0.71
M	0.55	0.45-0.65	0.63	0.55-0.71	0.78
OS	0.75	0.72-0.77	0.80	0.78-0.82	0.90
Overall	N/A	N/A	N/A	N/A	0.94

4. Discussion

This study aimed to develop a comprehensive evaluation scale for LLM-powered counseling chatbots (CES-LCC) leveraging domain knowledge of a pool of experts using the eDelphi method. Through two rounds of expert feedback, the scale was refined to address a broad spectrum of aspects related to the evaluation of this type of counseling chatbots. The results, particularly the qualitative feedback obtained using open ended questions, highlight the importance of a multidisciplinary approach to developing tools that effectively evaluate the different aspects and functionalities of modern digital health solutions. The final version of CES-LCC includes 27 items across nine dimensions and offers a robust framework to assess the unique challenges and capabilities of LLM-powered counseling chatbots. The structured eDelphi process facilitated the identification and integration of critical evaluation dimensions, leading to significant refinements. For instance, the addition of a "Memory" dimension underscores the need to assess chatbots' abilities to retain and build upon previous interactions, a functionality critical for creating a coherent and personalized user experience with digital mental health interventions. The importance of this dimension is grounded in scientific literature in the concepts of Memory Support Intervention [66], [67] in which information of previous sessions is embedded by the therapist in the ongoing dialogue, session summaries, and skill-building exercises to enhance retention, facilitate continuity, and promote the practical application of therapeutic concepts in the patient's daily life. The exclusion of a "Privacy and security" dimension from the scale instead reflects the need for a focused approach to evaluation in which chatbots and their infrastructural aspects (e.g. production choices like selecting encryption methods, implementing security protocols, adhering to privacy regulations) are evaluated separately. This is in line with ISO/IEC 25010, which distinguishes between different quality characteristics in system and security evaluation. In this standard, security (encompassing attributes such as confidentiality, integrity, and authenticity) is treated as a distinct nonfunctional requirement, separate from usability or functional suitability [68]. The expert panel also emphasized the importance of avoiding anthropomorphic language when assessing trust for avoiding misattributing human-like qualities to AI agents. This insight reflects the broader challenge related to both the design of transparent AI agents and the frameworks to assess such systems. Anthropomorphism is often used to increase retention and to promote self-disclosure [69], however using it in the context of mental health poses a great risk of exacerbating maladaptive behaviors and thoughts (e.g. social isolation) [70]. As a result, evaluation methods must address the unique capabilities of AI-driven systems while avoiding the promotion of anthropomorphic views, particularly in contexts where such perspectives could pose significant risks.

 doi:10.20944/preprints202501.1621.v1

4.1. Implications

By addressing gaps in existing methodologies, CES-LCC offers a framework for comprehensively assessing LLM-powered counseling chatbots. Unlike traditional evaluation tools that often focus on single aspects, this scale captures a broader spectrum of dimensions, including emotional support, memory retention, and trustworthiness. This study highlights the potential value and utility of the developed evaluation scale, although its applicability and impact require further validation and exploration. For researchers, the scale provides an integrated tool that can facilitate systematic investigations into the effectiveness of counseling chatbots. By combining technical and relational dimensions, the scale encourages multidisciplinary studies, potentially fostering deeper insights into how these technologies interact with users in complex, emotionally charged scenarios. This could contribute to the development of more sophisticated chatbot designs and the refinement of LLM technologies in therapeutic settings. In practice, the scale may serve as a useful tool for developers, mental health practitioners, and policymakers to evaluate and improve counseling chatbots. Developers could use the scale to identify specific areas for enhancement, ensuring their chatbots meet the demands of the users. Mental health practitioners might find the scale helpful when selecting chatbots to integrate into their services, as it provides a structured way to assess their potential. Finally, policymakers, particularly those involved in healthcare technology regulation, could leverage the scale to establish benchmarks for chatbot performance and safety.

4.2. Limitations and Future Research

While this study provides valuable insights, several limitations must be acknowledged. First of all, the experts sample size was limited, and because the participant pool included only experts from Italy, the generalizability of the findings to broader cultural or professional contexts may be limited. Further validation efforts must address these geographical and demographic limitations by incorporating international perspectives. Additionally, although an initial real-world validation was conducted, it involved filling CES-LCC after a single session delivered by a chatbot to a limited group of users (n=49), which restricts the robustness of the conclusions regarding the scale practical application and reliability. Comprehensive real-world testing with diverse user groups is needed to assess the scale's reliability and utility across various scenarios. The scale's development is still in its nascent stages, and its psychometric properties require further investigation. Finally, the reliance on the eDelphi method, which depends on subjective expert judgment, introduces potential biases despite efforts to ensure diverse expertise and minimize influence of individual perspectives. Future iterations should aim to integrate additional methodologies (e.g. factor analysis) to corroborate and enhance the objectivity of the findings.

5. Conclusions

This study presents the CES-LCC, a comprehensive evaluation scale developed to assess the unique challenges posed by evaluating LLM-powered counseling chatbots. Through an iterative eDelphi process involving multidisciplinary experts, the scale captures critical dimensions such as emotional support, trust, memory retention, and overall satisfaction. Initial validation in a real-world setting indicates strong reliability, emphasizing its potential utility for researchers, developers, and practitioners. The scale's multidimensional approach encourages a holistic assessment of chatbot performance, facilitating the identification of areas for enhancement. Despite its limitations, including the reliance on a geographically restricted expert panel for its development and limited user validation, the CES-LCC represents a significant step forward in standardizing the evaluation of modern counseling chatbots. Future research should focus on broader validation efforts, integrating diverse user perspectives, exploring the scale's psychometric properties, and examining its applicability in real-world contexts more extensively.

 doi:10.20944/preprints202501.1621.v1

Author Contributions: M.B. and S.G. contributed substantially to the conception and design of the study, to the acquisition of data and to the editing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Hub Life Science – Digital Health (LSH-DH) PNC-E3-2022-23683267 - Project DHEAL-COM - CUP C63C22001970001, Ministry of Health (Italy) under the Piano Nazionale Complementare al PNRR Ecosistema Innovativo della Salute (Code PNC-E.3).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is available on request.

Acknowledgments: During the preparation of this work the authors used ChatGPT (GPT-4o) to improve readability and language. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Conflicts of Interest: The authors declare no conflicts of interest.

 doi:10.20944/preprints202501.1621.v1

Appendix A

Appendix A.1. Round 1 Results Overview (Items)

Dim	Item	Relevance				Priority			Sample QL Feedback
		% Agree	M (SD)	Mdn (R)	IQR	M (SD)	Mdn (R)	IQR	
UR	The chatbot consistently understands what I am saying or asking.	100.00	4.88 (0.34)	5 (4-5)	0.00	4.88 (0.34)	5 (4-5)	0.00	*- I'd add another item about the tone (e.g. The chatbot understands the tone of my request)* *- I believe an important feature is not only that it understands but also infers from the sentence*
	I have to rephrase my requests very often for the chatbot to understand.	62.50	3.69 (1.14)	4 (1-5)	1.25	3.88 (0.89)	4 (3-5)	2.00	*N/A*
PHI	The chatbot provides accurate and helpful information.	87.50	4.44 (1.09)	5 (1-5)	1.00	4.38 (1.15)	5 (1-5)	1.00	*- I would divide the question into two separate questions. Answers might be accurate, but not helpful and vice versa, so 1) the chatbot provides accurate information; 2) the chatbot provides helpful information* *- I think it's important to add that the information it provides is grounded in theoretical frameworks and scientific literature.*
	The chatbot often provides incorrect or incomplete information.	75.00	3.94(1.18)	4 (1-5)	1.25	3.69 (1.25)	4 (1-5)	0.75	*N/A*
CRR	The chatbot's responses are clear, concise, and easy to understand.	81.25	4.06 (1.00)	4 (2-5)	1.00	4.00 (1.03)	4 (2-5)	1.25	*- Conciseness and easiness are very different dimensions, I don't feel like they should be evaluated together* *- I'd add an item about verbosity (e.g. The chatbot adds superfluous information related to the query)*
	The chatbot's responses are confusing or irrelevant to my questions.	81.25	4.00 (1.32)	4 (1-5)	1.00	3.94 (1.12)	4 (1-5)	1.00	*- I'd split the item in two (confusing/irrelevant)*
EUI	The chatbot is easy to interact with.	62.50	3.94 (1.00)	4 (2-5)	2.00	3.63 (1.31)	3.5 (1-5)	2.00	*- I have no better way. My doubt is about the definition of easy. What does it mean? How can someone evaluate this dimension?*
	Using the chatbot is frustrating or requires too many tentative interactions.	62.50	3.81 (1.11)	4 (1-5)	2.00	3.56 (1.21)	3.5 (2-5)	2.25	*- The item is not very clear; does it refer to access or the actual internal use of the LLM?*
LQ	The chatbot uses correct grammar and spelling in its responses.	75.00	3.88 (1.02)	4 (1-5)	0.50	3.50 (1.21)	4 (1-5)	1.00	*N/A*
	The chatbot's language style is natural and appropriate for the context.	68.75	4.00 (0.97)	4 (2-5)	2.00	3.50 (1.37)	3.5 (1-5)	2.00	*- I would divide this item into 2 items: 1) The chatbot's language style is/seems natural; 2) The chatbot's language is appropriate for the context.*
T	I believe the chatbot has my best interests at heart.	43.75	3.13 (1.50)	3 (1-5)	2.25	3.13 (1.50)	3 (1-5)	2.25	*- I would avoid formulations that suggest that the chatbot might have a cognition.* *- I believe that it is tricky to talk about the chatbot as if it has agency and conscience also because it might lead to overreliance and/or excessive anthropomorphization of the tool.*
	I am willing to rely on the chatbot in the future.	62.50	3.81 (0.91)	4 (2-5)	1.25	3.50 (1.15)	3.5 (2-5)	1.50	*N/A*
ES	The chatbot makes me feel heard and understood.	75.00	4.06 (1.12)	4 (1-5)	1.25	3.94 (1.18)	4 (1-5)	1.25	*- I'd add an item about sense of humor (e.g. The chatbot shows to have a sense of humor when required)*
	The chatbot's responses feel empathetic and supportive.	87.50	4.19 (0.98)	4 (2-5)	1.00	4.00 (0.89)	4 (1-5)	0.00	*- I'd add an item about feeling reassured (e.g. The chatbot's inputs and responses can make me feel reassured)*
GD	The chatbot provides helpful advice and suggestions for coping with my problems.	93.75	4.38 (0.62)	4 (3-5)	1.00	4.25 (0.68)	4 (3-5)	1.00	*- This question might crossload into the "Providing helpful information" factor. However, I would keep it, because in this case it specifically talks about coping, but maybe phrasing it as follows:*

 doi:10.20944/preprints202501.1621.v1

	The chatbot encourages me to take positive steps towards my goals.	75.00	3.88 (0.81)	4 (2-5)	0.25	3.81 (0.83)	4 (2-5)	1.00	*The chatbot provides adjusted guidance in coping with my problems.* *- I believe it is important to assess an individual's goals carefully. For example, a person with an eating disorder might set a goal to lose an extreme amount of weight, which is unhealthy. Therefore, it's crucial to remember that a patient's goals are not always the best for their well-being.*
OS	I am overall satisfied with the usability of this chatbot.	87.50	4.44 (0.73)	5 (3-5)	1.00	4.44 (0.73)	5 (3-5)	1.00	*- I want to point out that not only the usability but also the effectiveness in helping is important.*
	I would not recommend this chatbot to others due to usability issues.	75.00	3.88 (1.15)	4 (1-5)	1.25	3.88 (1.31)	4 (1-5)	2.00	*- This statement sounds somehow redundant to the first one in terms that lower scores on statement 1 seem to be almost equivalent to high*

Appendix A.2. Round 1 Qualitative Feedback (General and New Dimensions)

Feedback type	Content
New dimensions	*- I think assessing memory quality is crucial when dealing with real world implementations using LLMs. So, I suggest adding this dimension to the assessment.* *- I believe there are missing items related to the perception of how the chatbot handles my privacy and data security, such as how it shares my personal information with third parties.* *- In my opinion, privacy and data security, especially regarding how the chatbot shares personal information, are important aspects not included in the items. However, these seem more tied to production or implementation and might be better addressed in a separate, dedicated evaluation.* *- Items related to data privacy and security might be relevant in this scenario. However, in my experience, these items are more aligned with production or implementation processes and might be better addressed under regulations like the EU AI Act and GDPR.*
General	*- Almost all the statements sound actually very relevant, I provided some lower scores in some of them just to distinguish the ones I think are most relevant but in general all are relevant!* *- I felt like all of the shown questions were relevant in some way: that's why some evaluations were a bit harsh, just so that I could express what is more relevant from my point of view. Anyway, the questions were all pretty clear* *- An important consideration for real-life implementation of LLM-powered chatbots is ensuring accessibility for a wide range of users. This includes compatibility with various devices, such as smartphones, tablets, and computers, to meet diverse user needs. Additionally, designing the chatbot to be inclusive is crucial—for example, allowing users to specify preferred names and pronouns to support transgender and gender-diverse individuals, and incorporating features like colorblind-friendly graphics or text presentation options to assist users with visual impairments or reading difficulties. These steps can significantly enhance user experience and inclusivity.* *- Psychometrically, factors should have at least 3 items to be considered reliable, with 2 items it is not even possible to calculate internal consistency*

Appendix A.3. Round 1 decision

Dim	Add (Motivation)	Modify (Motivation)	Drop (Motivation)
UR	- "The chatbot understands the tone of my request" (QL Feedback) - "The chatbot asks specific questions to better understand my requests" (QL Feedback) - "The chatbot infers information from my messages" (QL Feedback)	Nothing	- "I have to rephrase my requests very often for the chatbot to understand." (% Agree Relevance)
PHI	- "The chatbot provides information grounded in theory and scientific literature." (QL Feedback) - "The chatbot provides references." (QL Feedback)	- Split the item "The chatbot provides accurate and helpful information." into "The chatbot provides accurate information" and "The chatbot provides helpful information" (QL Feedback) - Split the item "The chatbot often provides incorrect or incomplete information." into "The chatbot often provides incorrect information." and "The chatbot often provides incomplete information." (QL Feedback)	Nothing
CRR	- "The chatbot adds superfluous information related to the query" (QL Feedback)	- Split the item "The chatbot's responses are clear, concise, and easy to understand." into two different items "The chatbot's responses are clear, and easy to understand", "The chatbot's responses are adequately concise" (QL Feedback)	Nothing

 doi:10.20944/preprints202501.1621.v1

		- Split the item "The chatbot's responses are confusing or irrelevant to my questions." into "The chatbot's responses are confusing" and "The chatbot's responses are irrelevant to my questions." (QL Feedback)	
EUI	Nothing	Nothing	- The entire dimension (% Agree Relevance, IQR Relevance, QL Feedback)
LQ	Nothing	- Split the item "The chatbot's language style is natural and appropriate for the context." into "The chatbot's language style is/seems natural" and "The chatbot's language is appropriate for the context." (% Agree Relevance, IQR Relevance, QL Feedback)	Nothing
T	- "I feel safe sharing my personal matters with the chatbot" (QL Feedback) - "I believe that the feedback/information provided by the chatbot are trustworthy" (QL Feedback) - "I believe the chatbot is transparent about its limitations and capabilities." (QL Feedback)	Nothing	- "I believe the chatbot has my best interests at heart" (% Agree Relevance, IQR Relevance, QL Feedback) - "I am willing to rely on the chatbot in the future." (% Agree Relevance)
ES	- "The chatbot's responses can make me feel reassured" (QL Feedback) - "The chatbot shows to have a sense of humor when required" (QL Feedback)	Nothing	Nothing
GD	- "The chatbot helps me set realistic and achievable goals." (QL Feedback)	- Modify "The chatbot provides helpful advice and suggestions for coping with my problems." into "The chatbot provides adjusted guidance in coping with my problems" to avoid cross loading with other factor (QL Feedback) - Modify "The chatbot encourages me to take positive steps." (QL Feedback)	Nothing
OS	- "I am overall satisfied with the effectiveness of this chatbot" (QL Feedback) - "I feel that my interactions with the chatbot were worthwhile." (QL Feedback)	Nothing	- "I would not recommend this chatbot to others due to usability issues." (Redundancy, QL Feedback)
M [New]	- "The chatbot accurately recalls key details from previous conversations." (QL Feedback) - "The chatbot maintains consistency by integrating past interactions into current responses." (QL Feedback) - "The chatbot adapts its advice based on information provided in earlier sessions." (QL Feedback)	Nothing	Nothing

Appendix B

Appendix B.1. Round 2 Results Overview (Items)

Dim	Item (Italian)	Relevance				Redund	Priority	Translation Qual		Sample QL Feedback
		% Agree	M (SD)	Mdn (R)	IQR	Flags	Points	% Agree	M (SD)	
UR	The chatbot consistently understands what I am saying or asking. (Il chatbot capisce sempre ciò che sto dicendo o chiedendo.)	100.00	4.87 (0.35)	5 (4-5)	0.00	1	56	73.33	4.73 (0.47)	*- I'd prefer "The chatbot consistently understands what I am saying AND asking". The "or" makes it hard to trust high scores. [Content]* *- Toglierei il "sempre" che in italiano potrebbe inserire un dubbio invece che rafforzare [Translation]*

 doi:10.20944/preprints202501.1621.v1

	The chatbot understands the tone of my request. (Il chatbot capisce il tono della mia richiesta.)	73.33	4.07 (0.80)	4 (3-5)	1.50	3	27	73.33	4.67 (0.65)	*- It is not clear to me what "understanding the tone" means here [Content]*
	The chatbot asks specific questions to better understand my requests. (Il chatbot fa domande specifiche per capire meglio le mie richieste.)	86.67	4.20 (0.68)	4 (3-5)	1.00	0	30	80.00	4.92 (0.29)	*N/A*
	The chatbot infers information from my messages. (Il chatbot inferisce informazioni dai miei messaggi.)	80.00	4.43 (0.94)	5 (2-5)	1.00	2	37	60.00	4.25 (1.06)	*- The term "infer" is a bit ambiguous; I would suggest revising it as follows: "The chatbot is able to make adequate inferences based on my messages." [Content]* *- Il chatbot deduce informazioni dai miei messaggi [Translation]*
PHI	The chatbot provides accurate information. (Il chatbot fornisce informazioni accurate.)	86.67	4.47 (1.09)	5 (2-5)	1.00	2	81	80.00	4.83 (0.39)	*N/A*
	The chatbot provides helpful information. (Il chatbot fornisce informazioni utili.)	100.00	4.80 (0.41)	5 (4-5)	0.00	0	68	80.00	4.92 (0.29)	*N/A*
	The chatbot often provides incorrect information. (Il chatbot fornisce spesso informazioni errate.)	80.00	4.13 (1.13)	4 (1-5)	1.00	4	52	80.00	4.92 (0.29)	*N/A*
	The chatbot often provides incomplete information. (Il chatbot fornisce spesso informazioni incomplete.)	73.33	4.00 (0.76)	4 (3-5)	1.00	1	53	80.00	4.92 (0.29)	*N/A*
	The chatbot provides information grounded in theory and scientific literature. (Il chatbot fornisce informazioni basate su teorie e letteratura scientifica.)	80.00	4.13 (1.13)	4 (1-5)	1.00	3	35	73.33	4.50 (0.67)	*- Il chatbot fornisce informazioni supportate da teorie e letteratura [Translation]*
	The chatbot provides references. (Il chatbot fornisce riferimenti bibliografici.)	66.67	3.60 (1.06)	4 (1-5)	1.00	5	26	60.00	4.42 (0.90)	*- I don't think it's crucial for users to have a research paper attached to questions such as "I feel bad lately, I can't sleep". It would make the UX poorer in my opinion. This would make more sense if you are building a search engine kind of system. [Content]* *- Il chatbot fornisce riferimenti alle fonti utilizzate [Translation]*
CRR	The chatbot's responses are clear, and easy to understand. (Le risposte del chatbot sono chiare e facili da capire.)	100.00	4.93 (0.26)	5 (4-5)	0.00	0	57	73.33	4.75 (0.62)	*- Le risposte del chatbot sono chiare e semplici da capire [Translation]*
	The chatbot's responses are adequately concise. (Le risposte del chatbot sono sufficientemente concise.)	80.00	4.13 (0.74)	4 (3-5)	1.00	1	53	80.00	4.75 (0.45)	*N/A*
	The chatbot's responses are confusing. (Le risposte del chatbot sono confondenti.)	93.33	4.07 (0.96)	4 (1-5)	0.50	5	50	53.33	3.83 (1.19)	*- This is just the reverse of clear [Content]* *- Le risposte del chat mi confondono [Translation]*
	The chatbot's responses are irrelevant to my questions. (Le risposte del chatbot non sono pertinenti alle mie domande.)	100.00	4.73 (0.46)	5 (4-5)	0.50	1	42	80.00	4.92 (0.29)	*N/A*
	The chatbot adds superfluous information related to the query. (Il chatbot aggiunge informazioni superflue relative alla richiesta.)	53.33	3.47 (0.92)	4 (1-5)	1.00	8	23	60.00	4.45 (1.29)	*- Il chatbot aggiunge informazioni superflue rispetto alla richiesta. [Translation]*
LQ	The chatbot uses correct grammar and spelling in its responses. (Il chatbot fornisce risposte con grammatica e ortografia corrette.)	80.00	3.73 (1.22)	4 (1-5)	0.00	2	34	60.00	4.42 (0.90)	*- Il chatbot fornisce risposte grammaticalmente e ortograficamente corrette. [Translation]*
	The chatbot's language style is/seems natural. (Lo stile linguistico del chatbot è/sembra naturale.)	86.67	4.47 (0.74)	5 (3-5)	1.00	0	23	66.67	4.67 (0.78)	*- "The chatbot's language style sounds natural" seems more fluent [Content]* *- Lo stile linguistico del chatbot suona naturale [Translation]*
	The chatbot's language is appropriate for the context. (Il linguaggio del chatbot è appropriato per il contesto.)	86.67	4.57 (0.65)	5 (3-5)	1.00	0	33	73.33	4.58 (0.67)	*- Il linguaggio del chatbot è appropriato al contesto. [Translation]*
T	I feel safe sharing my personal matters with the chatbot. (Mi sento al sicuro nel condividere questioni personali con il chatbot.)	93.33	4.53 (1.06)	5 (1-5)	0.50	0	37	80.00	4.75 (0.45)	*N/A*
	I believe the chatbot is transparent about its limitations and capabilities. (Credo che il chatbot sia trasparente riguardo alle sue limitazioni e capacità.)	73.33	4.13 (0.83)	4 (3-5)	1.50	0	21	53.33	4.18 (1.08)	*- Credo che il chatbot sia trasparente riguardo ai suoi limiti e alle sue capacità [Translation]*
	I believe that the feedback/information provided by the chatbot are trustworthy. (Credo che i feedback/le informazioni fornite dal chatbot siano affidabili.)	93.33	4.67 (0.82)	5 (2-5)	0.00	2	32	66.67	4.90 (0.32)	*- Mettere una e invece che la slash / [Translation]*
ES	The chatbot makes me feel heard and understood. (Il chatbot mi fa sentire ascoltato e capito.)	86.67	4.20 (1.08)	4 (1-5)	1.00	1	53	66.67	4.80 (0.42)	*- I would drop this. I feel like this evaluates how the system can trick the user in terms of feeling like they are talking to someone that*

 doi:10.20944/preprints202501.1621.v1

										listens to them and understands, while an LLM obviously cannot do that. [Content]
	The chatbot's responses feel empathetic and supportive. (Le risposte del chatbot sembrano empatiche e di supporto.)	93.33	4.60 (0.63)	5 (3-5)	1.00	1	43	53.33	4.10 (1.29)	*- I think this is different to the previous one, because it focuses on the "look" of the answers more than on the ability of convincing the user of something. This is something that makes sense to evaluate I think [Content]* *- "e supportive" invece che "di supporto" [Translation]*
	The chatbot's responses can make me feel reassured (Le risposte del chatbot sono in grado di farmi sentire rassicurato.)	80.00	4.20 (0.77)	4 (3-5)	1.00	3	34	60.00	4.70 (0.67)	*N/A*
	The chatbot shows to have a sense of humor when required (Il chatbot dimostra di avere senso dell'umorismo quando necessario.)	60.00	3.20 (1.42)	4 (1-5)	2.00	4	20	60.00	4.70 (0.67)	*N/A*
GD	The chatbot provides adjusted guidance in coping with my problems. (Il chatbot mi fornisce indicazioni adeguate per affrontare i problemi che riporto.)	86.67	4.53 (0.74)	5 (3-5)	1.00	0	38	46.67	3.90 (1.29)	*- Nel tradurre coping suggerirei di dire "per gestire" invece che per affrontare [Translation]* *- Il chatbot fornisce indicazioni adeguate per affrontare i miei problem [Translation]*
	The chatbot helps me set realistic and achievable goals. (Il chatbot mi aiuta a stabilire obiettivi realistici e raggiungibili.)	100.00	4.53 (0.52)	5 (4-5)	1.00	1	25	66.67	4.80 (0.42)	*N/A*
	The chatbot encourages me to take positive steps. (Il chatbot mi incoraggia a compiere sforzi per il mio benessere.)	86.67	4.27 (1.03)	5 (2-5)	1.00	2	27	53.33	3.82 (0.87)	*- Il chatbot mi incoraggia a compiere azioni costruttive. [Translation]* *- Il chatbot mi incoraggia a compiere passi positivi [Translation]*
	The chatbot accurately recalls key details from previous conversations. (Il chatbot ricorda accuratamente i dettagli chiave delle conversazioni precedenti.)	100.00	4.73 (0.46)	5 (4-5)	0.50	1	39	60.00	4.55 (0.82)	*- I would delate "key", to not make it seem like if the chatbot can understand personal saliance, but rather its capacity to recall information at large this item is important. [Content]*
M	The chatbot maintains consistency by integrating past interactions into current responses. (Il chatbot integra coerentemente le interazioni passate nelle risposte attuali.)	93.33	4.80 (0.56)	5 (3-5)	0.00	4	26	53.33	4.50 (0.85)	*- Il chatbot è coerente ed integra le interazioni passate nelle risposte attuali. [Translation]* *- Il chatbot integra coerentemente le interazioni passate nelle risposte [Translation]*
	The chatbot adapts its advice based on information provided in earlier sessions. (Il chatbot adatta i suoi consigli in base alle informazioni fornite nelle sessioni precedenti.)	93.33	4.67 (0.62)	5 (3-5)	0.50	6	25	73.33	4.82 (0.40)	*N/A*
	I am overall satisfied with the usability of this chatbot. (Sono complessivamente soddisfatto dell'usabilità di questo chatbot.)	93.33	4.53 (0.64)	5 (3-5)	1.00	0	39	73.33	4.64 (0.50)	*- Nel complesso, sono soddisfatto dell'usabilità di questo chatbot [Translation]*
OS	I feel that my interactions with the chatbot were worthwhile. (Trovo che le mie interazioni con il chatbot siano state utili.)	86.67	4.20 (0.68)	4 (3-5)	1.00	3	27	66.67	4.73 (0.65)	*- Trovo che le mie interazioni con il chatbot siano state proficue [Translation]*
	I am overall satisfied with the effectiveness of this chatbot. (Sono complessivamente soddisfatto dell'efficacia di questo chatbot.)	73.33	4.20 (1.01)	5 (2-5)	1.50	2	24	60.00	4.70 (0.67)	*- Nel complesso, sono soddisfatto... [Translation]*

Appendix B.2. Round 2 Decision

Dim	Add (Motivation)	Modify (Motivation)	Drop (Motivation)
UR	Nothing	- Rephrase both the italian translation and the original item ("The chatbot consistently understands what I am saying or asking.") into: "The chatbot consistently understands what I am saying and asking" and "Il chatbot capisce ciò che sto dicendo e chiedendo." (% Agree Translation, QL Feedback) - Rephrase both the italian translation and the original item ("The chatbot infers information from my messages.") into: "The chatbot is able to make adequate inferences based on my messages." and "Il chatbot è in grado di fare deduzioni appropriate basandosi sui miei messaggi." (% Agree Translation, QL Feedback)	- "The chatbot understands the tone of my request." (% Agree Relevance, QL Feedback)
PHI	Nothing	- Rephrase the italian version of the item "The chatbot provides information grounded in theory and scientific literature." into "Il chatbot fornisce informazioni supportate da teorie e letteratura scientifica." (% Agree Translation, QL Feedback)	- The chatbot often provides incorrect information. (Redundancy) - The chatbot often provides incomplete information. (% Agree Relevance) - The chatbot provides references. (% Agree Relevance, Redundancy)

 doi:10.20944/preprints202501.1621.v1

CRR	Nothing	- Rephrase the italian version of the item “The chatbot's responses are clear, and easy to understand.” into “Le risposte del chatbot sono chiare e semplici da capire.” (% Agree Translation, QL Feedback)	- The chatbot's responses are confusing. (Redundancy, QL Feedback) - The chatbot adds superfluous information related to the query. (% Agree Relevance, Redundancy, QL Feedback)
LQ	Nothing	- Rephrase the italian version of the item “The chatbot uses correct grammar and spelling in its responses.” into “Il chatbot fornisce risposte grammaticalmente e ortograficamente corrette.” (% Agree Translation, QL Feedback) - Rephrase both the italian translation and the original item (“The chatbot's language style is/seems natural.”) into: “The chatbot's language style sounds natural.” and “Lo stile linguistico del chatbot suona naturale.” (% Agree Translation, QL Feedback) - Rephrase the italian version of the item “The chatbot's language is appropriate for the context.” into “Il linguaggio del chatbot è appropriato al contesto.” (% Agree Translation, QL Feedback)	Nothing
T	Nothing	- Rephrase the italian version of the item “I believe the chatbot is transparent about its limitations and capabilities.” into “Credo che il chatbot sia trasparente riguardo ai suoi limiti e alle sue capacità” (% Agree Translation, QL Feedback) - Rephrase both the italian translation and the original item (“I believe that the feedback/information provided by the chatbot are trustworthy.”) into: “I believe that the feedback and the information provided by the chatbot are trustworthy.” and “Credo che i feedback e le informazioni fornite dal chatbot siano affidabili.” (% Agree Translation, QL Feedback)	Nothing
ES	Nothing	- Rephrase the italian version of the item “The chatbot's responses feel empathetic and supportive.” into “Le risposte del chatbot risultano empatiche e supportive.” (% Agree Translation, QL Feedback)	- “The chatbot shows to have a sense of humor when required” (% Agree Relevance, Redundancy, QL Feedback)
GD	Nothing	- Rephrase the italian version of the item “The chatbot provides adjusted guidance in coping with my problems.” into “Il chatbot fornisce indicazioni personalizzate per aiutarmi a gestire i miei problemi.” (% Agree Translation, QL Feedback) - Rephrase the italian version of the item “The chatbot encourages me to take positive steps.” into “Il chatbot mi incoraggia a compiere azioni costruttive.” (% Agree Translation, QL Feedback)	Nothing
M	Nothing	- Rephrase both the italian translation and the original item (“The chatbot accurately recalls key details from previous conversations.”) into: “The chatbot accurately recalls details from previous conversations.” and “Il chatbot ricorda accuratamente i dettagli delle conversazioni precedenti.” (% Agree Translation, QL Feedback) - Rephrase the italian version of the item “The chatbot maintains consistency by integrating past interactions into current responses.” into “Il chatbot integra coerentemente le interazioni passate nelle risposte.” (% Agree Translation, QL Feedback)	Nothing
OS	Nothing	- Rephrase the italian version of the item “I am overall satisfied with the usability of this chatbot.” into Nel complesso, sono soddisfatto dell'usabilità di questo chatbot.” (% Agree Translation, QL Feedback) - Rephrase both the italian translation and the original item (“I feel that my interactions with the chatbot were worthwhile.”) into: “Overall, I feel that my interactions with the chatbot were worthwhile.” and “Nel complesso, trovo che le mie interazioni con il chatbot siano state proficue.” (% Agree Translation, QL Feedback) - Rephrase both the italian translation and the original item (“I am overall satisfied with the effectiveness of this chatbot.”) into: “I am overall satisfied with the support provided by this chatbot.” and “Nel complesso, sono soddisfatto del supporto offerto da questo chatbot.” (% Agree Translation, % Agree Relevance, QL Feedback)	Nothing

 doi:10.20944/preprints202501.1621.v1

Appendix C

Appendix C.1. Demographic Profile of Users who Participated in the Initial Validation

Characteristic		Value or % (n)
Age		M = 32.02 (SD = 11.55)
Gender	Female	57.14% (28)
	Male	40.81% (20)
	Not specified	2.05% (1)
Education	EQF1	0.00% (0)
	EQF2	8.16% (4)
	EQF3	2.04% (1)
	EQF4	14.29% (7)
	EQF5	0.00% (0)
	EQF6	28.57% (14)
	EQF7	30.61% (15)
	EQF8	16.33% (8)
Chatbot Experience	None	18.37% (9)
	Basic	32.65% (16)
	Intermediate	34.69% (17)
	Expert	14.29% (7)
LLM Experience	None	24.49% (12)
	Basic	38.78% (19)
	Intermediate	26.53% (13)
	Expert	10.20% (5)
Propensity to Trust in Technology [71]		M = 3.76 (SD = 0.51)
Country	Italy	100% (49)

References

1. E. Bendig, B. Erb, L. Schulze-Thuesing, and H. Baumeister, "The Next Generation: Chatbots in Clinical Psychology and Psychotherapy to Foster Mental Health – A Scoping Review," *Verhaltenstherapie*, vol. 32, no. Suppl. 1, pp. 64–76, 2022, doi: 10.1159/000501812.
2. M. Laymouna, Y. Ma, D. Lessard, T. Schuster, K. Engler, and B. Lebouché, "Roles, Users, Benefits, and Limitations of Chatbots in Health Care: Rapid Review," *J Med Internet Res*, vol. 26, p. e56930, Jul. 2024, doi: 10.2196/56930.
3. L. Balcombe, "AI Chatbots in Digital Mental Health," *Informatics*, vol. 10, no. 4, p. 82, Oct. 2023, doi: 10.3390/informatics10040082.
4. E. M. Boucher *et al.*, "Artificially intelligent chatbots in digital mental health interventions: a review," *Expert Rev Med Devices*, vol. 18, no. sup1, pp. 37–49, Dec. 2021, doi: 10.1080/17434440.2021.2013200.
5. M. Skjuve, A. Følstad, and P. B. Brandtzaeg, "The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users," in *Proceedings of the 5th International Conference on Conversational User Interfaces*, New York, NY, USA: ACM, Jul. 2023, pp. 1–10. doi: 10.1145/3571884.3597144.
6. S. Limpanopparat, E. Gibson, and D. A. Harris, "User engagement, attitudes, and the effectiveness of chatbots as a mental health intervention: A systematic review," *Computers in Human Behavior: Artificial Humans*, vol. 2, no. 2, p. 100081, Aug. 2024, doi: 10.1016/j.chbah.2024.100081.
7. H. L. O'Brien and E. G. Toms, "What is user engagement? A conceptual framework for defining user engagement with technology," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 6, pp. 938–955, Apr. 2008, doi: 10.1002/asi.20801.
8. M. Hassenzahl and N. Tractinsky, "User experience - a research agenda," *Behaviour & Information Technology*, vol. 25, no. 2, pp. 91–97, Mar. 2006, doi: 10.1080/01449290500330331.
9. B. Shackel, "Usability – Context, framework, definition, design and evaluation," *Interact Comput*, vol. 21, no. 5–6, pp. 339–346, Dec. 2009, doi: 10.1016/j.intcom.2009.04.007.
10. J. Moilanen, A. Visuri, S. A. Suryanarayana, A. Alorwu, K. Yatani, and S. Hosio, "Measuring the Effect of Mental Health Chatbot Personality on User Engagement," in *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia*, New York, NY, USA: ACM, Nov. 2022, pp. 138–150. doi: 10.1145/3568444.3568464.

 doi:10.20944/preprints202501.1621.v1

11. S. Gabrielli *et al.*, "Engagement and Effectiveness of a Healthy-Coping Intervention via Chatbot for University Students During the COVID-19 Pandemic: Mixed Methods Proof-of-Concept Study," *JMIR Mhealth Uhealth*, vol. 9, no. 5, p. e27965, May 2021, doi: 10.2196/27965.
12. H. L. O'Brien and E. G. Toms, "The development and evaluation of a survey to measure user engagement," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 50–69, Jan. 2010, doi: 10.1002/asi.21229.
13. K. Denecke, S. Vaaheesan, and A. Arulnathan, "A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test," *IEEE Trans Emerg Top Comput*, vol. 9, no. 3, pp. 1170–1182, Jul. 2021, doi: 10.1109/TETC.2020.2974478.
14. C. G. Escobar-Viera, G. Porta, R. W. S. Coulter, J. Martina, J. Goldbach, and B. L. Rollman, "A chatbot-delivered intervention for optimizing social media use and reducing perceived isolation among rural-living LGBTQ+ youth: Development, acceptability, usability, satisfaction, and utility," *Internet Interv*, vol. 34, p. 100668, Dec. 2023, doi: 10.1016/j.invent.2023.100668.
15. M. R. Lima, M. Wairagkar, N. Natarajan, S. Vaitheswaran, and R. Vaidyanathan, "Robotic Telemedicine for Mental Health: A Multimodal Approach to Improve Human-Robot Engagement," *Front Robot AI*, vol. 8, Mar. 2021, doi: 10.3389/frobt.2021.618866.
16. B. Laugwitz, T. Held, and M. Schrepp, "Construction and Evaluation of a User Experience Questionnaire," 2008, pp. 63–76. doi: 10.1007/978-3-540-89350-9_6.
17. J. Shah *et al.*, "Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening," *International Journal of Eating Disorders*, vol. 55, no. 9, pp. 1229–1244, Sep. 2022, doi: 10.1002/eat.23798.
18. K. Boyd *et al.*, "Usability testing and trust analysis of a mental health and wellbeing chatbot," in *Proceedings of the 33rd European Conference on Cognitive Ergonomics*, New York, NY, USA: ACM, Oct. 2022, pp. 1–8. doi: 10.1145/3552327.3552348.
19. M. N. Islam, S. R. Khan, N. N. Islam, Md. Rezwan-A-Rownok, S. R. Zaman, and S. R. Zaman, "A Mobile Application for Mental Health Care During COVID-19 Pandemic: Development and Usability Evaluation with System Usability Scale," 2021, pp. 33–42. doi: 10.1007/978-3-030-68133-3_4.
20. S. Valtolina, P. Zanotti, and S. Mandelli, "Designing Conversational Agents to Empower Active Aging," in *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, New York, NY, USA: ACM, Sep. 2024, pp. 1–4. doi: 10.1145/3652988.3673939.
21. J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," in *Usability Evaluation In Industry*, CRC Press, 1996, pp. 207–212. doi: 10.1201/9781498710411-35.
22. S. Holmes, A. Moorhead, R. Bond, H. Zheng, V. Coates, and M. Mctear, "Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces?," in *Proceedings of the 31st European Conference on Cognitive Ergonomics*, New York, NY, USA: ACM, Sep. 2019, pp. 207–214. doi: 10.1145/3335082.3335094.
23. S. Borsci *et al.*, "The Chatbot Usability Scale: the Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents," *Pers Ubiquitous Comput*, vol. 26, no. 1, pp. 95–119, Feb. 2022, doi: 10.1007/s00779-021-01582-9.
24. T. Henkel, A. J. Linn, and M. J. van der Goot, "Understanding the Intention to Use Mental Health Chatbots Among LGBTQIA+ Individuals: Testing and Extending the UTAUT," 2023, pp. 83–100. doi: 10.1007/978-3-031-25581-6_6.
25. T. Kamita, T. Ito, A. Matsumoto, T. Munakata, and T. Inoue, "A Chatbot System for Mental Healthcare Based on SAT Counseling Method," *Mobile Information Systems*, vol. 2019, pp. 1–11, Mar. 2019, doi: 10.1155/2019/9517321.
26. Venkatesh, Morris, Davis, and Davis, "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, vol. 27, no. 3, p. 425, 2003, doi: 10.2307/30036540.
27. F. D. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly*, vol. 13, no. 3, p. 319, Sep. 1989, doi: 10.2307/249008.

28. K. Ahuja and P. Lio, "Measuring Empathy in Artificial Intelligence: Insights From Psychodermatology and Implications for General Practice," *Prim Care Companion CNS Disord*, vol. 26, no. 5, Oct. 2024, doi: 10.4088/PCC.24lr03782.
29. J. Zhao, F. M. Plaza-del-Arco, B. Genchel, and A. C. Curry, "Language Model Council: Democratically Benchmarking Foundation Models on Highly Subjective Tasks," Jun. 2024.
30. M. Schmidmaier, J. Rupp, D. Cvetanova, and S. Mayer, "Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: ACM, May 2024, pp. 1–18. doi: 10.1145/3613904.3642035.
31. S. Concannon and M. Tomalin, "Measuring perceived empathy in dialogue systems," *AI Soc*, vol. 39, no. 5, pp. 2233–2247, Oct. 2024, doi: 10.1007/s00146-023-01715-z.
32. A. Miloff, P. Carlbring, W. Hamilton, G. Andersson, L. Reuterskiöld, and P. Lindner, "Measuring Alliance Toward Embodied Virtual Therapists in the Era of Automated Treatments With the Virtual Therapist Alliance Scale (VTAS): Development and Psychometric Evaluation," *J Med Internet Res*, vol. 22, no. 3, p. e16660, Mar. 2020, doi: 10.2196/16660.
33. S. Wei, D. Freeman, and A. Rovira, "A randomised controlled test of emotional attributes of a virtual coach within a virtual reality (VR) mental health treatment," *Sci Rep*, vol. 13, no. 1, p. 11517, Jul. 2023, doi: 10.1038/s41598-023-38499-7.
34. H. Q. Yu and S. McGuinness, "An experimental study of integrating fine-tuned large language models and prompts for enhancing mental health support chatbot system," *J Med Artif Intell*, vol. 7, pp. 16–16, Jun. 2024, doi: 10.21037/jmai-23-136.
35. R. Crasto, L. Dias, D. Miranda, and D. Kayande, "CareBot: A Mental Health ChatBot," in *2021 2nd International Conference for Emerging Technology (INCET)*, IEEE, May 2021, pp. 1–5. doi: 10.1109/INCET51464.2021.9456326.
36. A. Srivastava, I. Pandey, M. S. Akhtar, and T. Chakraborty, "Response-act Guided Reinforced Dialogue Generation for Mental Health Counseling," in *Proceedings of the ACM Web Conference 2023*, New York, NY, USA: ACM, Apr. 2023, pp. 1118–1129. doi: 10.1145/3543507.3583380.
37. M. N. Kaysar and S. Shiramatsu, "Mental State-Based Dialogue System for Mental Health Care by Using GPT-3," 2024, pp. 891–901. doi: 10.1007/978-981-99-3043-2_74.
38. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.
39. C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out*, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.
40. N. M. Radziwill and M. C. Benton, "Evaluating Quality of Chatbots and Intelligent Conversational Agents," Apr. 2017.
41. H. Ding, J. Simmich, A. Vaezipour, N. Andrews, and T. Russell, "Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review," *Journal of the American Medical Informatics Association*, vol. 31, no. 3, pp. 746–761, Feb. 2024, doi: 10.1093/jamia/ocad222.
42. H. Donohoe, M. Stellefson, and B. Tennant, "Advantages and Limitations of the e-Delphi Technique," *Am J Health Educ*, vol. 43, no. 1, pp. 38–46, Jan. 2012, doi: 10.1080/19325037.2012.10599216.
43. I. Belton, A. MacDonald, G. Wright, and I. Hamlin, "Improving the practical application of the Delphi method in group-based judgment: A six-step prescription for a well-founded and defensible process," *Technol Forecast Soc Change*, vol. 147, pp. 72–82, Oct. 2019, doi: 10.1016/j.techfore.2019.07.002.
44. S. S. McMillan, M. King, and M. P. Tully, "How to use the nominal group and Delphi techniques," *Int J Clin Pharm*, Feb. 2016, doi: 10.1007/s11096-016-0257-x.
45. S. Jünger, S. A. Payne, J. Brine, L. Radbruch, and S. G. Brearley, "Guidance on Conducting and REporting DElphi Studies (CREDES) in palliative care: Recommendations based on a methodological systematic review," *Palliat Med*, vol. 31, no. 8, pp. 684–706, Sep. 2017, doi: 10.1177/0269216317690685.
46. K. Denecke, R. May, and O. Rivera Romero, "Potential of Large Language Models in Health Care: Delphi Study," *J Med Internet Res*, vol. 26, p. e52399, May 2024, doi: 10.2196/52399.

47. W. Maroengsit, T. Piyakulpinyo, K. Phonyiam, S. Pongnumkul, P. Chaovalit, and T. Theeramunkong, "A Survey on Evaluation Methods for Chatbots," in *Proceedings of the 2019 7th International Conference on Information and Education Technology*, New York, NY, USA: ACM, Mar. 2019, pp. 111–119. doi: 10.1145/3323771.3323824.
48. K. Denecke, A. Abd-Alrazaq, M. Househ, and J. Warren, "Evaluation Metrics for Health Chatbots: A Delphi Study," *Methods Inf Med*, vol. 60, no. 05/06, pp. 171–179, Dec. 2021, doi: 10.1055/s-0041-1736664.
49. Z. Guo, A. Lai, J. H. Thygesen, J. Farrington, T. Keen, and K. Li, "Large Language Model for Mental Health: A Systematic Review," Feb. 2024, doi: 10.2196/preprints.57400.
50. T. Y. C. Tam *et al.*, "A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review," May 2024.
51. Y. Chang *et al.*, "A Survey on Evaluation of Large Language Models," *ACM Trans Intell Syst Technol*, vol. 15, no. 3, pp. 1–45, Jun. 2024, doi: 10.1145/3641289.
52. J.-L. Peng *et al.*, "A Survey of Useful LLM Evaluation," Jun. 2024.
53. "Qualtrics XM," 2024.
54. "Mistral Large 2407," 2024.
55. D. G. Saari, "Selecting a voting method: the case for the Borda count," *Constitutional Political Economy*, vol. 34, no. 3, pp. 357–366, Sep. 2023, doi: 10.1007/s10602-022-09380-y.
56. V. Clarke and V. Braun, "Thematic analysis," *J Posit Psychol*, vol. 12, no. 3, pp. 297–298, May 2017, doi: 10.1080/17439760.2016.1262613.
57. M. Sheinis and A. Selk, "Development of the Adult Vulvar Lichen Sclerosus Severity Scale—A Delphi Consensus Exercise for Item Generation," *J Low Genit Tract Dis*, vol. 22, no. 1, pp. 66–73, Jan. 2018, doi: 10.1097/LGT.0000000000000361.
58. S. M. Bauer, A. Fusté, A. Andrés, and C. Saldaña, "The Barcelona Orthorexia Scale (BOS): development process using the Delphi method," *Eating and Weight Disorders - Studies on Anorexia, Bulimia and Obesity*, vol. 24, no. 2, pp. 247–255, Apr. 2019, doi: 10.1007/s40519-018-0556-4.
59. T. Xin, X. Ding, H. Gao, C. Li, Y. Jiang, and X. Chen, "Using Delphi method to develop Chinese women's cervical cancer screening intention scale based on planned behavior theory," *BMC Womens Health*, vol. 22, no. 1, p. 512, Dec. 2022, doi: 10.1186/s12905-022-02113-1.
60. V. C. Scott, J. Temple, and Z. Jillani, "Development of the Technical Assistance Engagement Scale: a modified Delphi study," *Implement Sci Commun*, vol. 5, no. 1, p. 84, Jul. 2024, doi: 10.1186/s43058-024-00618-4.
61. World Health Organization, *Doing What Matters in Times of Stress: An Illustrated Guide*. 2020.
62. L. J. Cronbach, "Coefficient Alpha and the Internal Structure of Tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951, doi: 10.1007/BF02310555.
63. J. P. Guilford, "The Correlation of an Item With a Composite of the Remaining Items in a Test," *Educ Psychol Meas*, vol. 13, no. 1, pp. 87–93, Apr. 1953, doi: 10.1177/001316445301300109.
64. M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *Int J Med Educ*, vol. 2, pp. 53–55, Jun. 2011, doi: 10.5116/ijme.4dfb.8dfd.
65. A. Röschel, C. Wagner, and M. Dür, "Examination of validity, reliability, and interpretability of a self-reported questionnaire on Occupational Balance in Informal Caregivers (OBI-Care) – A Rasch analysis," *PLoS One*, vol. 16, no. 12, p. e0261815, Dec. 2021, doi: 10.1371/journal.pone.0261815.
66. G. G. Zieve, L. D. Sarfan, L. Dong, S. S. Tiab, M. Tran, and A. G. Harvey, "Cognitive Therapy-as-Usual versus Cognitive Therapy plus the Memory Support Intervention for adults with depression: 12-month outcomes and opportunities for improved efficacy in a secondary analysis of a randomized controlled trial," *Behaviour Research and Therapy*, vol. 170, p. 104419, Nov. 2023, doi: 10.1016/j.brat.2023.104419.
67. L. Dong *et al.*, "Can integrating the Memory Support Intervention into cognitive therapy improve depression outcome? A randomized controlled trial," *Behaviour Research and Therapy*, vol. 157, p. 104167, Oct. 2022, doi: 10.1016/j.brat.2022.104167.
68. S. Ouhbi, A. Idri, J. L. Fernández-Alemán, A. Toval, and H. Benjelloun, "Applying ISO/IEC 25010 on Mobile Personal Health Records," in *Proceedings of the International Conference on Health Informatics*, SCITEPRESS - Science and and Technology Publications, 2015, pp. 405–412. doi: 10.5220/0005216604050412.

 doi:10.20944/preprints202501.1621.v1

69. M. Blut, C. Wang, N. V. Wünderlich, and C. Brock, "Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI," *J Acad Mark Sci*, vol. 49, no. 4, pp. 632–658, Jul. 2021, doi: 10.1007/s11747-020-00762-y.
70. F. Eyssel and N. Reich, "Loneliness makes the heart grow fonder (of robots) — On the effects of loneliness on psychological anthropomorphism," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, IEEE, Mar. 2013, pp. 121–122. doi: 10.1109/HRI.2013.6483531.
71. S. Jessup, T. Schneider, G. Alarcon, T. Ryan, and A. Capiola, "The Measurement of the Propensity to Trust Technology," 2018.